Winnow Assignment:

1. Convert the dataset into a csv file

2. I have coded in Spark using Python. I have decided to use Naive Bayes algorithm for classification of whether it is a phishing website or not.

Code:

phishNB\_RDD = sc.textFile("file:///home/cloudera/Downloads/PhishNB.txt")

from pyspark.mllib.regression import LabeledPoint

def newrow(line):

outrow = []

list = line.split(",")

for x in xrange(30):

outrow.append(list[x])

return (LabeledPoint(list[30],outrow))

intNB\_RDD = phishNB\_RDD.map(newrow)

from pyspark.mllib.classification import NaiveBayes

my\_nbmodel = NaiveBayes.train(intNB\_RDD)

print my\_nbmodel

datax\_col=intNB\_RDD.collect()

trainset\_pred =[]

for x in datax\_col:

trainset\_pred.append(my\_nbmodel.predict(x.features))

print trainset\_pred

nb\_cf\_mat=np.zeros([2,2]) #num of classes

for pnt in datax\_col:

predctn = my\_nbmodel.predict(np.array(pnt.features))

nb\_cf\_mat[pnt.label][predctn]+=1

corrcnt=0

for i in range(2):

corrcnt+=nb\_cf\_mat[i][i]

nb\_per\_corr=corrcnt/nb\_cf\_mat.sum()

print 'Naive Bayes: Conf.Mat. and Per Corr'

print nb\_cf\_mat

print nb\_per\_corr

Output for sample weather dataset with 7 variabes(sunny, overcast, rainy, temp, humid, rainy, result-play) and 14 instances:

1.0,0.0,0.0,85.0,85.0,0.0,0.0

1.0,0.0,0.0,80.0,90.0,1.0,0.0

0.0,1.0,0.0,83.0,86.0,0.0,1.0

0.0,0.0,1.0,70.0,96.0,0.0,1.0

0.0,0.0,1.0,68.0,80.0,0.0,1.0

0.0,0.0,1.0,65.0,70.0,1.0,0.0

0.0,1.0,0.0,64.0,65.0,1.0,1.0

1.0,0.0,0.0,72.0,95.0,0.0,0.0

1.0,0.0,0.0,69.0,70.0,0.0,1.0

0.0,0.0,1.0,75.0,80.0,0.0,1.0

1.0,0.0,0.0,75.0,70.0,1.0,1.0

0.0,1.0,0.0,72.0,90.0,1.0,1.0

0.0,1.0,0.0,81.0,75.0,0.0,1.0

0.0,0.0,1.0,71.0,91.0,1.0,0.0

Naive Bayes: Conf.Mat. and Per Corr

[[ 3. 2.]

[ 0. 9.]]

0.857142857143